Facultad de Ingeniería

ISO 19157 - Información geográfica - Calidad de datos

Hebenor Bermúdez - Miguel Gavirondo

Octubre 2023

FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Descriptores de los elementos

La evaluación de un elemento se describe con:

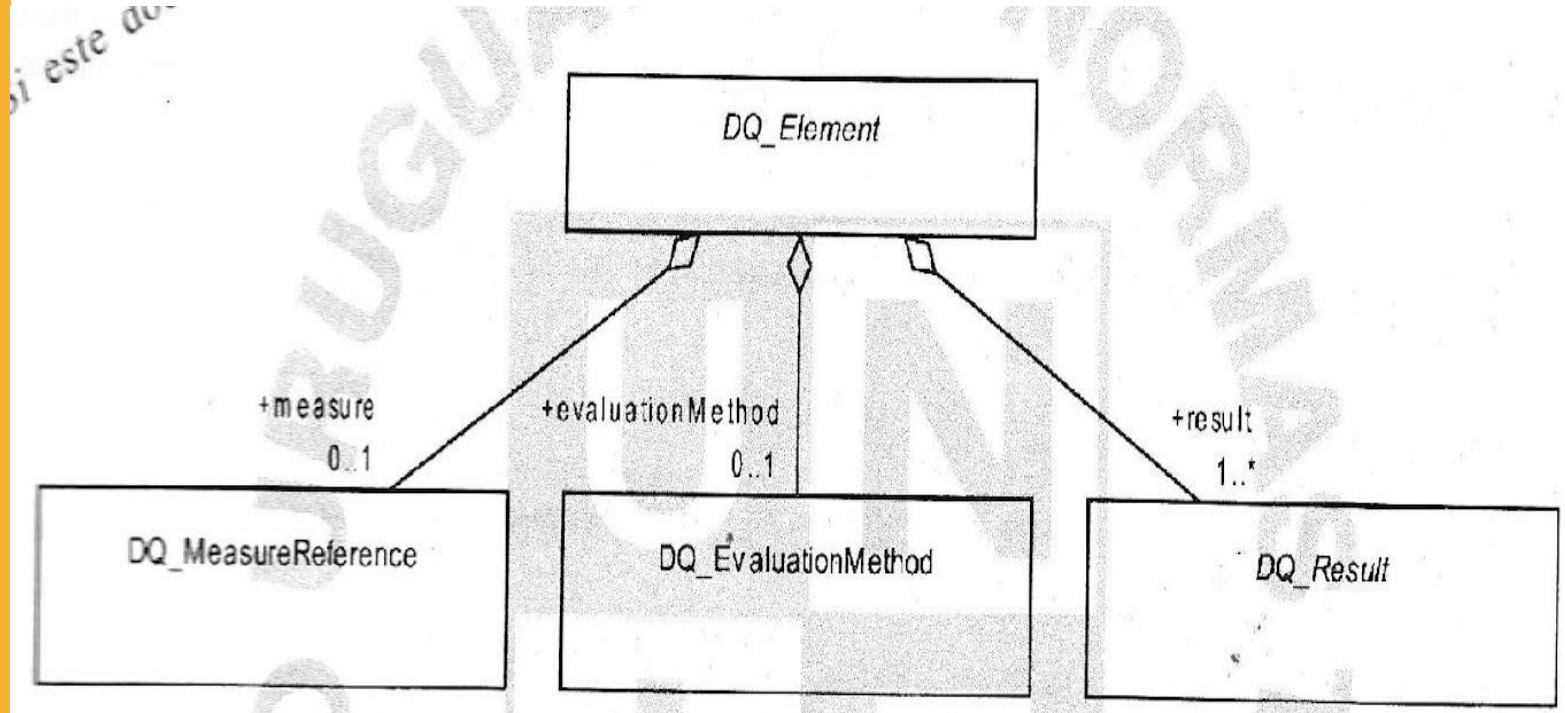


Figura 5 – Descriptores de un elemento de la calidad de datos

Medidas

Las medidas de la calidad es la forma que tenemos de medir la calidad de un conjunto de datos geográficos en cada uno de los elementos de la calidad.

El objetivo es que las distintas evaluaciones sean comparables entre ellas de manera independiente de su fuente.

Para eso deben usarse las medidas normalizadas de la calidad que figuran en la norma (Anexo D) siempre que sea posible ("...se recomienda encarecidamente…").

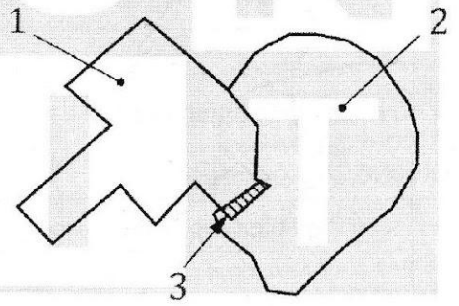
Medidas

Las medidas normalizadas de la calidad que se brindan abarcan todos los elementos de la calidad y se da más de una medida para cada uno de ellos.

CATEGORÍA	ELEMENTO	N.º DE MEDIDAS (Anexo D)
COMPLECIÓN	Comisión	4
	Omisión	4
CONSISTENCIA LÓGICA	Consistencia conceptual	6
	Consistencia de dominio	5
	Consistencia de formato	3
	Consistencia topológica	7
EXACTITUD POSICIONAL	Exactitud absoluta	5
	Exactitud relativa	2
	Exactitud relativa	2
	Exactitud posicional de datos en malla	(*)
CALIDAD TEMPORAL	Exactitud en la medida del tiempo	6
	Consistencia temporal	1
	Validez temporal	(**)
EXACTITUD TEMÁTICA	Corrección de la clasificación	5
	Corrección de atributo no cuantitativo	3
	Corrección de atributo cuantitativo	6
	USABILIDAD	5

Medidas

Tabla D.11 – Número de superposiciones no válidas entre superficies

Línea	Componente	Descripción
1	Nombre	número de superposiciones no válidas entre superficies
2	Alias	superficies superpuestas
3	Nombre del elemento	consistencia conceptual
4	Medida básica	recuento de errores
5	Definición	número total de superposiciones erróneas en los datos
6	Descripción	depende de la aplicación cuáles superficies pueden y cuáles no deben superponerse. No todas las superficies superpuestas son erróneas. Al informar sobre esta medida de la calidad de datos, debe informarse también sobre el tipo de clases de objeto geográfico de las superficies ilegalmente superpuestas
7	Parámetro	–
8	Tipo de valor	entero
9	Estructura del valor	–
10	Fuente de referencia	–
11	Ejemplo	 Leyenda 1 Superficie 1 2 Superficie 2 3 Área superpuesta
12	Identificador	11

Medidas

Tabla D.32 – Número de incertidumbres posicionales mayores que un umbral

Línea	Componente	Descripción
1	Nombre	número de incertidumbres posicionales mayores que un umbral
2	Alias	–
3	Nombre del elemento	exactitud absoluta o externa
4	Medida básica	recuento de errores
5	Definición	número de incertidumbres posicionales superiores a un umbral dado para un conjunto de posiciones los errores se definen como la distancia entre la posición medida y la que se considera como verdadera
6	Descripción	para un número de puntos (N), se ofrecen las posiciones medidas como coordenadas x_{mi}, y_{mi} y z_{mi} dependiendo de las dimensiones en las que se mide la posición del punto. Se considera que un conjunto correspondiente de coordenadas, x_{ti}, y_{ti} y z_{ti}, representa las posiciones verdaderas. El cálculo de e_i se define en la medida de la calidad "valor medio de las incertidumbres posicionales" en una, dos y tres dimensiones se consideran como error todas las incertidumbres posicionales por encima del umbral predefinido $e_{máx.}$ ($e_i > e_{máx.}$) debería fijarse un criterio para el establecimiento de correspondencias (por ejemplo, permitiendo considerar vértices a lo largo de líneas para correspondencias con la posición más cercana). En el resultado de evaluación de la calidad se debe informar sobre el/los criterio/s para encontrar los puntos homólogos
7	Parámetro	nombre: $e_{máx.}$ definición: umbral de aceptación de incertidumbres posicionales tipo de valor: numérico
8	Tipo de valor	entero
9	Estructura del valor	–

Medidas

Tabla D.59 – Exactitud temporal al 95% de nivel de significación

Línea	Componente	Descripción
1	Nombre	exactitud temporal al 95% de nivel de significación
2	Alias	–
3	Nombre del elemento	exactitud de una medida de tiempo
4	Medida básica	LE95 o LE95(r), dependiendo del procedimiento de evaluación
5	Definición	mitad de la longitud del intervalo, definido por un límite superior y otro inferior, en que se sitúa el valor verdadero de la instancia de tiempo con una probabilidad del 95%
6	Descripción	véase G.3.2
7	Parámetro	–
8	Tipo de valor	medida
9	Estructura del valor	–
10	Fuente de referencia	–
11	Ejemplo	–
12	Identificador	57

Medidas

Tabla D.65 – Matriz de error de la clasificación

Línea	Componente	Descripción
1	Nombre	matriz de error de la clasificación
2	Alias	matriz de confusión
3	Nombre del elemento	corrección de la clasificación
4	Medida básica	
5	Definición	matriz que indica el número de ítems de la clase (i) clasificados como clase (j)
6	Descripción	la matriz de error de la clasificación (MCM, misclassification matrix) es una matriz cuadrada de *n* columnas y *n* filas. *n* indica el número de clases consideradas MCM (i,j) = [nº de ítems de la clase (i) clasificados como clase (j)] los elementos de la diagonal de la matriz de error de clasificación contiene los ítems clasificados correctamente, y los elementos fuera de la diagonal contienen el número de errores de la clasificación
7	Parámetro	nombre: n definición: número de clases consideradas tipo de valor: entero
8	Tipo de valor	entero
9	Estructura del valor	matriz $(n \times n)$
10	Fuente de referencia	–
11	Ejemplo	
12	Identificador	62

		Clase del conjunto de datos			
		A	B	C	Recuento
Clase verdadera	A	7	2	1	10
	B	1	2	2	5
	C	1	1	3	5
	Recuento	9	5	6	20

Medidas

Tabla D.80 – Índice de especificaciones de producto no satisfechas

Línea	Componente	Descripción
1	Nombre	índice de especificaciones de producto no satisfechas
2	Alias	–
3	Nombre del elemento	usabilidad
4	Medida básica	índice de error
5	Definición	número de requisitos de las especificaciones de producto de datos que han sido incumplidos en el producto/conjunto de datos actual en relación al número total de requisitos
6	Descripción	–
7	Parámetro	–
8	Tipo de valor	real
9	Estructura del valor	–
10	Fuente de referencia	–
11	Ejemplo	–
12	Identificador	104

Medidas

Aparte de las medidas normalizadas los usuarios **pueden crear** sus propias medidas.

Antes de crear una nueva medida, siempre hay que asegurarse que esa medida **ya no exista** en la norma y que la calidad no puede medirse con alguna de las medidas normalizadas.

Para describir una nueva medida se puede recurrir a las medidas básicas (ver Anexo G) y utilizando la estructura establecida en la norma.

Nuevas medidas
Estructura

Para crear una nueva medida se deben completar los siguientes puntos:

1) **IDENTIFICADOR DE LA MEDIDA** (o): Valor que identifica de manera única una medida dentro de un espacio de nombres.
2) **NOMBRE** (o): nombre de la medida.
3) **ALIAS** (op): nombre alternativo de la medida (otro nombre, abreviatura, nombre corto). Puede tener más de uno.
4) **NOMBRE DEL ELEMENTO** (o): nombre del elemento de la calidad al que se aplica la medida.
5) **MEDIDA BÁSICA** (c): si la medida se basa en una medida básica (Anexo G) se debe indicar el nombre de la misma.

Nuevas medidas
Estructura

Para crear una nueva medida se deben completar los siguientes puntos:

6) DEFINICIÓN (o): es la parte fundamental de la medida.
7) DESCRIPCIÓN (c): se debe incluir los métodos de cálculo, fórmulas, y las ilustraciones que se entiendan necesarias para determinar correctamente el resultado de la medida. Si se trabaja con el concepto de error hay que dejar claro cuando una unidad se considera errónea.
8) PARÁMETRO (c): variable auxiliar utilizada en la medida. Se debe incluir nombre, definición y tipo de valor (ver tablas D.65, D.66 y D.67)

Nuevas medidas
Estructura

Para crear una nueva medida se deben completar los siguientes puntos:

9) TIPO DE VALOR (o): tipo de dato usado para reportar el resultado. Están definidos por la norma ISO 19103. Ej: entero, real, booleano, matriz, etc.
10) ESTRUCTURA DEL VALOR (op): en el caso de que el resultado sea más de un valor el mismo debe estructurarse.
11) FUENTE DE REFERENCIA (c): si la medida procede de una fuente externa (otra organización, otro documento) se debe hacer referencia a la misma.
12) EJEMPLO (op): se puede proporcionar uno o más ejemplos de la medida de forma de facilitar su comprensión.

Medidas básicas

La norma define **medidas básicas** ya que hay medidas de la calidad que tienen características comunes. Ej: el conteo de errores se puede usar para construir distintas medidas como porcentajes, índices, etc.

Las medidas básicas se pueden identificar en dos categorías con el recuento y la incertidumbre.

Las medidas de recuento se basan en contar errores o de elementos correctos.

Las medidas de incertidumbre se basan en el modelado de incertidumbre a través de métodos estadísticos.

Medidas básicas

Tabla G.1 – Medidas básicas de la calidad de datos para medidas de la calidad relacionadas con el recuento

Nombre de la medida básica	Definición de la medida básica	Ejemplo	Tipo de valor
Indicador de error	Indicador de que un ítem es incorrecto	Falso	Booleano (si el valor es verdadero el ítem es incorrecto)
Indicador de corrección	Indicador de que un ítem es correcto	Verdadero	Booleano (si el valor es verdadero el ítem es correcto)
Recuento de errores	Número total de ítems que poseen un error de una tipología concreta	11	Entero
Recuento de ítems correctos	Número total de ítems que están libres de errores de una tipología concreta	571	Entero
Índice de error	Número de ítems erróneos respecto al número total de ítems	0,0189	Real
Índice de ítems correctos	Número de ítems correctos respecto al número total de ítems	0,9811	Real

NOTA 1 El índice de error puede presentarse como porcentaje o como razón. Puede usarse la unidad del valor del resultado cuantitativo (véase 7.5.4.2) para especificar que el resultado se presenta como porcentaje o como razón.

NOTA 2 El índice de ítems correctos puede presentarse como porcentaje o como razón. Puede usarse la unidad del valor del resultado cuantitativo (véase 7.5.4.2) para especificar que el resultado se presenta como porcentaje o como razón.

Medidas básicas

Los valores numéricos solo pueden obtenerse con una exactitud determinada. Entonces la incertidumbre puede ser tratada como una variable aleatoria.

Los métodos estadísticos usados para la definición de las medidas básicas de la calidad relacionadas con la incertidumbre se basan en los siguientes supuestos:

- Las incertidumbres son homogéneas para todos los valores observados.
- No existe correlación entre los valores observados.
- Los valores observados siguen una distribución normal.

Medidas básicas

Tabla G.2 – Relación entre los cuantiles de la distribución normal y el nivel de significación

Probabilidad P	Cuantil	Medida básica	Nombre	Tipo de valor
$P = 50\%$	$u_{50\%}$ 0,6745	$u_{50\%} \cdot \sigma_Z$	LE50	Medida
$P = 68{,}3\%$	$u_{68{,}3\%} = 1$	$u_{68{,}3\%} \cdot \sigma_Z$	LE68.3	Medida
$P = 90\%$	$u_{90\%} = 1{,}645$	$u_{90\%} \cdot \sigma_Z$	LE90	Medida
$P = 95\%$	$u_{95\%} = 1{,}960$	$u_{95\%} \cdot \sigma_Z$	LE95	Medida
$P = 99\%$	$u_{99\%} = 2{,}576$	$u_{99\%} \cdot \sigma_Z$	LE99	Medida
$P = 99{,}8\%$	$u_{99{,}8\%} = 3$	$u_{99{,}8\%} \cdot \sigma_Z$	LE99.8	Medida

Ciclo de vida del dato

Los procesos de evaluación de la calidad se utilizan a lo largo de todo el ciclo de vida del producto. La **norma considera** que las fases de este ciclo son:

ESPECIFICACIONES DE PRODUCTO O REQUERIMIENTOS DE LOS USUARIOS: se pueden utilizar los procedimientos de evaluación para definir los niveles de conformidad del producto deseado. También se pueden especificar que procedimientos se deben aplicar para la evaluación.

Ciclo de vida del dato

La **norma considera** que las fases de este ciclo son:

PRODUCCIÓN: durante la producción se pueden aplicar métodos de evaluación que estén o no definidos en las especificaciones del producto. Cuando se utilizan durante la producción debería informarse en el linaje de los datos los procedimientos usados.

ENTREGA: se deben hacer una evaluación de la calidad para verificar el cumplimiento de las especificaciones del producto previo a la entrega de los datos del usuario. El conjunto de datos puede ser aceptado o rechazado en caso de ser rechazado una vez corregido deberá ser nuevamente evaluado.

Ciclo de vida del dato

La **norma considera** que las fases de este ciclo son:

USO: Cualquier usuario puede hacer las evaluaciones que entienda necesarias para comprobar que esos datos son adecuados para sus necesidades.

ACTUALIZACIÓN: los procedimientos para el control de la calidad se deben utilizar para controlar la calidad durante los procesos de actualización y para informar la calidad luego de finalizada la misma.

Proceso de evaluación

ESPECIFICAR LA(S) UNIDAD(ES) DE LA CALIDAD DE LOS DATOS: cada unidad se especifica por el ámbito y por los elementos de la calidad que se aplican.

ESPECIFICAR LA(S) MEDIDA(S) DE LA CALIDAD DE LOS DATOS: se deberá especificar al menos una medida para cada elemento de la calidad evaluado.

ESPECIFICAR LOS PROCEDIMIENTOS DE EVALUACIÓN DE LA CALIDAD DE LOS DATOS: aplicar uno o varios métodos de evaluación.

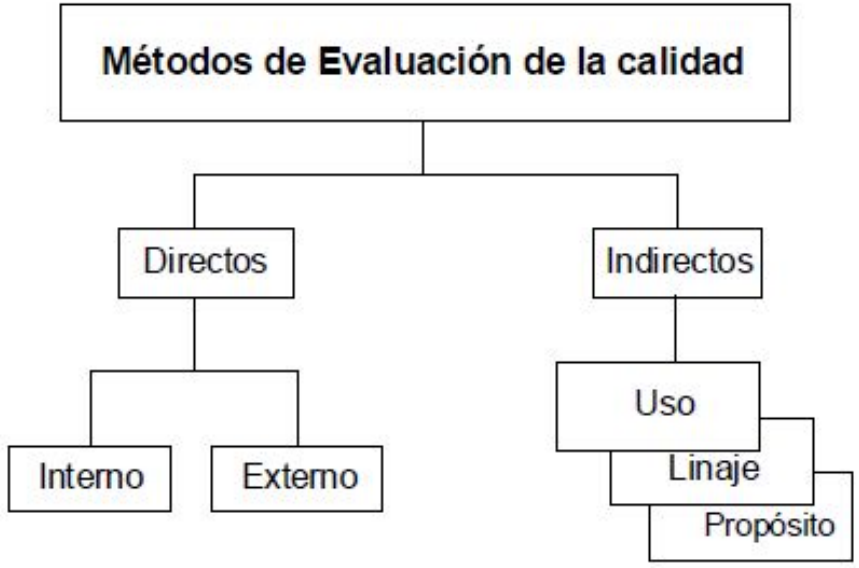
DETERMINAR LA SALIDA DE LA EVALUACIÓN DE LA CALIDAD: la salida de la aplicación de la evaluación es un resultado.

Métodos de evaluación

Los **métodos de evaluación** pueden dividirse en dos clases principales:

MÉTODOS DIRECTOS: utilizan información de referencia interna o externa para realizar la evaluación comparando esta información con los datos.

MÉTODOS INDIRECTOS: infieren o estiman la calidad usando información sobre los propios datos como ser el linaje, uso y propósito.



Métodos de evaluación

Los **MÉTODOS DIRECTOS** implican la inspección de los ítems del conjunto de datos.

Los **MÉTODOS DIRECTOS INTERNOS** Utilizan los datos que están en el propio conjunto de datos que se está evaluando.

Los **MÉTODOS DIRECTOS EXTERNOS** requiere de datos de referencia externos a los datos que se están evaluando.

Para evaluar la calidad por los métodos internos o externos se pueden usar la inspección total o el muestreo.

Métodos de evaluación

Los **MÉTODOS INDIRECTOS** se basan en el conocimiento o experiencia externa sobre los productos de datos. Puede ser **subjetivo**.

Para hacer esta evaluación se pueden utilizar fuentes **no cuantitativas** de información como el uso, linaje, propósito y otros documentos sobre los procedimientos de producción o sobre los insumos utilizados para producirlos (no se deben limitar a estos). También podría recurrirse a la **experiencia de expertos** en la temática.

Estos resultados suelen ser del tipo textual **sin resultados cuantitativos** ya que estos pueden ser engañosos. Requiere una buena documentación de cómo se hizo la evaluación.

Métodos de evaluación

A partir de los resultados de la calidad evaluados se pueden generar nuevos resultados sin realizar nuevas evaluaciones. Esto se puede realizar a través de la **AGREGACIÓN** y **DERIVACIÓN** de resultados.

La agregación **combina** resultados basados en diferentes elementos o ámbitos de la calidad.

La derivación **produce resultados** de calidad a partir de otros anteriores.

Informe de la calidad

Existen dos formas básicas de informar sobre la calidad:

- **a través de los metadatos.**
- **a través de informes independientes de la calidad.**

El informe de calidad a través de metadatos **ES OBLIGATORIO**. Siempre tiene que estar la calidad informada en los metadatos. Brinda información **breve, sintética y estructurada**.

El informe independiente de la calidad no tiene una estructura fija y se utiliza para dar **mayor detalles** sobre la calidad de los datos. Este informe **NO SUSTITUYE** los metadatos. Siempre debería usarse este informe cuando se obtienen resultados derivados y agregados de la calidad.

¿Por qué informar?

Debemos informar de la calidad por muchas razones, siendo las principales:

- Para ayudar a **encontrar** el conjunto de datos y fomentar su uso.
- Para **demostrar la conformidad** con unas especificaciones de producto o unos requisitos de usuario.
- Como parte de **iniciativas de gestión** de los proveedores.
- Para **permitir posteriores dictámenes** sobre la calidad de la información derivada del conjunto de datos.
- Para permitir la **toma racional de decisiones** cuando se sabe que todos los datos contienen defectos.

¿Por qué informar?

Los datos continuamente se están modificando (creando, actualizando, procesando, etc.) por lo que **la calidad varía** también **de forma continua**. Las tres circunstancia que pueden modificar la calidad de los datos son:

- Cuando se elimina, modifica o añade cualquier cantidad de datos al conjunto.
- Cuando se modifican las especificaciones del producto de datos o si se identifican nuevos requerimientos de los usuarios.
- Cuando cambia el mundo real.

Como se usan los elementos de la calidad

En algunos casos se pueden aplicar varios elementos de la calidad para comprobar una especificación.

COMPLECIÓN, **EXACTITUD TEMÁTICA** y **POSICIONAL** se utilizan para describir cómo se relaciona el conjunto de datos con el universo de discurso.

La **CONSISTENCIA LÓGICA** se puede evaluar sin conocer la realidad terreno. Evalúa las relaciones internas de los datos y el ajuste de los datos con las reglas establecidas en las especificaciones.

La **CALIDAD TEMPORAL** es una mezcla de las anteriores ya que algunas de ellas dependen de reglas lógicas y otras necesitan de la realidad terreno.

Como se usan los elementos de la calidad

En algunos casos se pueden aplicar varios elementos de la calidad para comprobar una especificación.

La **USABILIDAD** se puede utilizar para evaluar aspectos no contemplados en los elementos anteriores o para brindar resultados agregados a partir de los elementos anteriores o distintos ámbitos en los que se dividan los datos de evaluación.

Como se usan los elementos de la calidad

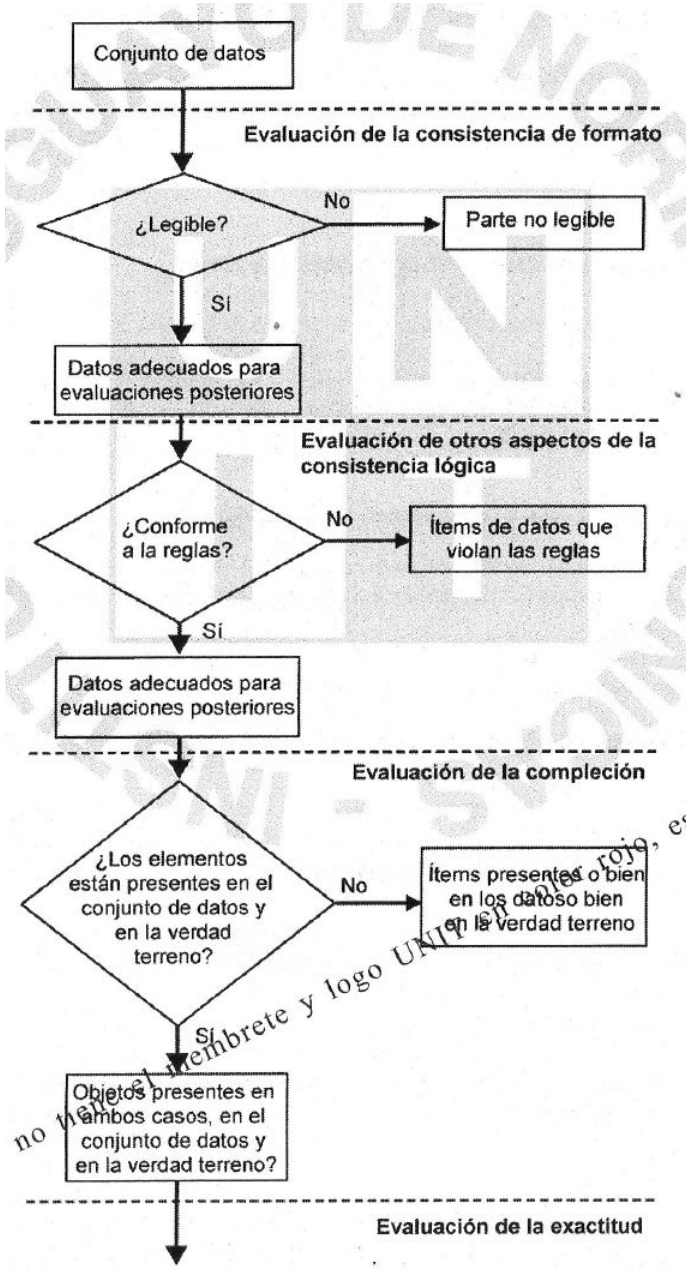


Figura I.1 – Orden en la evaluación de la calidad de datos

Precisión del control

Los procesos de control parten de la base de que los trabajos de observación **están libres de error**.

Sabiendo que eso no es así, se puede asumir que esta hipótesis es correcta si somos capaces de definir la **variable de interés** de forma **más precisa** que la de la **variable que se controla**.

¿Qué precisión se requiere en los trabajos de control respecto a los elementos que se controlan?

Comúnmente recurrimos a la regla de que **el método de relevamiento de la variable sea 3 veces más preciso**.

Precisión del control

Se se analiza por el lado de la composición de las varianzas se parte de lo siguiente:

$$\sigma_E = \sqrt{\sigma_P^2 + \sigma_C^2}$$

:

σ_E	Desviación de la estimación
σ_P	Desviación del producto
σ_C	Desviación del proceso de control

Si se parte de que la desviación del proceso de control es ⅓ de la del producto, haciendo las sustituciones respectivas se llega a que la desviación de la estimación es aproximadamente 1,05 de la desviación del producto.

En resumen la precisión de la estimación solo **aumenta** 5% a raíz del método de control.

Control posicional

En el control posicional tenemos las llamadas **Metodologías de Control Posicional por Puntos** (MCPP). Dentro de ellas encontramos los estándares más comunes como el NMAS, ASLSM o el NSSDA.

Para poder realizar estos procedimientos es necesario recurrir a una fuente de mayor exactitud la cual puede ser relevamiento directo o utilizando otro conjunto de datos geográficos. Por esto es necesario saber cual es la **exactitud esperada** para así poder definir el método de control.

Existen otros métodos de **control por elementos lineales** como por ejemplo banda épsilon, banda de error, etc.

Control posicional

Si no podemos conocer la exactitud esperada del producto:

- Se puede recurrir al criterio del límite de **apreciación visual** para fijar el **error radial máximo** del producto. Este error se asume como **estándar** así que hay que expandirlo a un nivel de confianza elevado, por ejemplo 95%.

- Las **técnicas GNSS** permiten metodologías de captura de datos fáciles y que aseguren altos niveles de exactitud.

Coherencia lógica

La coherencia lógica hace referencia al **grado de conformidad** de un conjunto de datos geográficos con **respecto a la estructura interna descripta en las especificaciones**.

Es de vital importancia en los productos digitales. Por eso es una componente que se controla de manera interna y que por lo tanto no requiere trabajo de campo aunque las rutinas de control pueden insumir un tiempo importante.

Al realizarse con rutinas automáticas se suelen realizar inspecciones al 100% o establecer NCA muy bajos.

Exactitud temática y compleción

Para evaluar estos elementos es necesario tener bien definido el **UNIVERSO DE DISCURSO**.

En el proceso de evaluación no se compara la cartografía con la realidad sino que debo hacerlo frente al terreno nominal, que se puede definir como el subconjunto de elementos del mundo real que cumplen con las especificaciones del usuario respecto del producto.

Lectura recomendada

- Anexo E partes 1, 2 y 3 de la norma
- Anexo I de la norma.